Performance Evaluation of Intel Nehalem Based Cluster

Subhash Saini

Subhash.Saini@nasa.gov

NASA Advanced Supercomputing Division (NAS)

NASA Ames Research Center

Moffett Field, California, USA

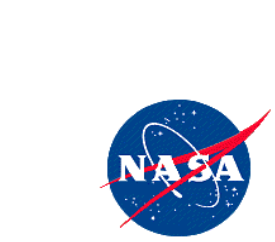
Los Alamos Computer Science Symposium (LACSS) 2009

Workshop on Performance Analysis of Extreme-Scale Systems and Applications

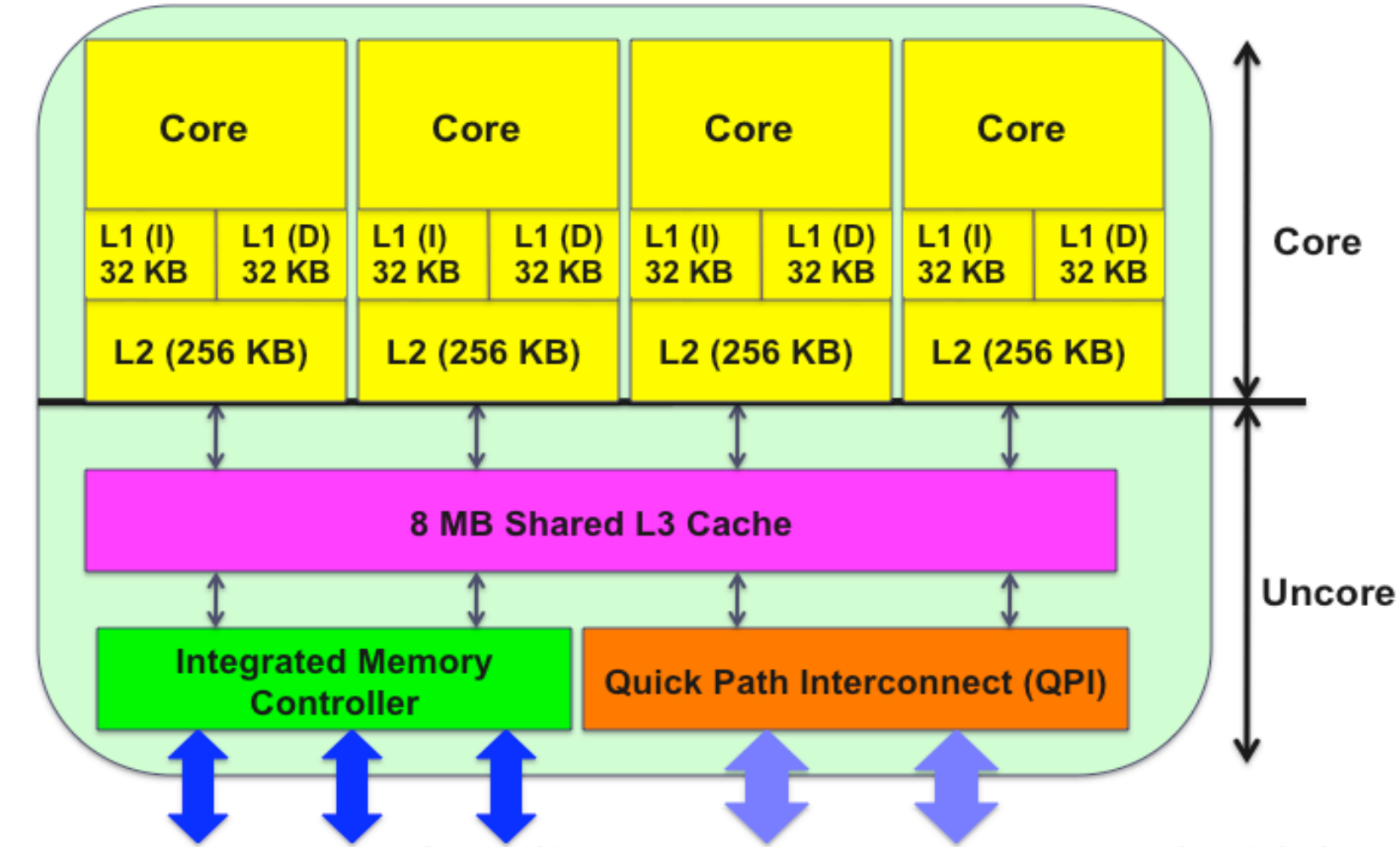
Santa Fe, New Mexico, October 13-14, 2009

Outline

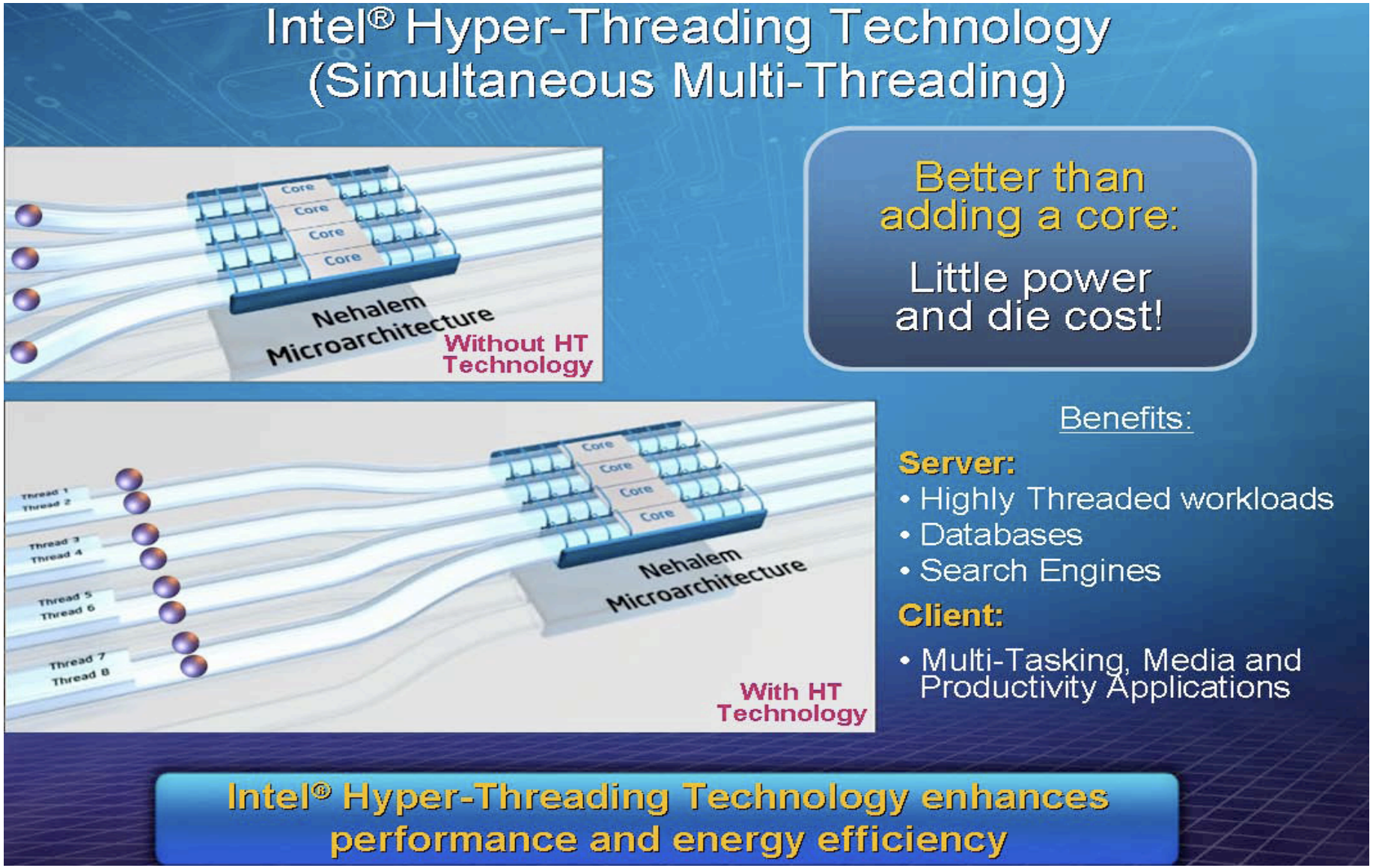
- Intel Nehalem cluster
  - Platform
    - Integrated memory controller
    - Quick Path Interconnect
    - DDR3 memory
  - CPU architecture
    - Intel Hyper-Threading technology
  - Power management
    - Power Gate
    - Turbo mode
  - Interconnect
    - Quad Data Rate (QDR) IB
- HPC challenge benchmarks (HPCC)
- Kernels & compact applications (NPB)
- Applications: Computational fluid dynamics, Climate, Molecular Dynamics and Earthquakes
- Conclusions



Intel Nehalem Processor

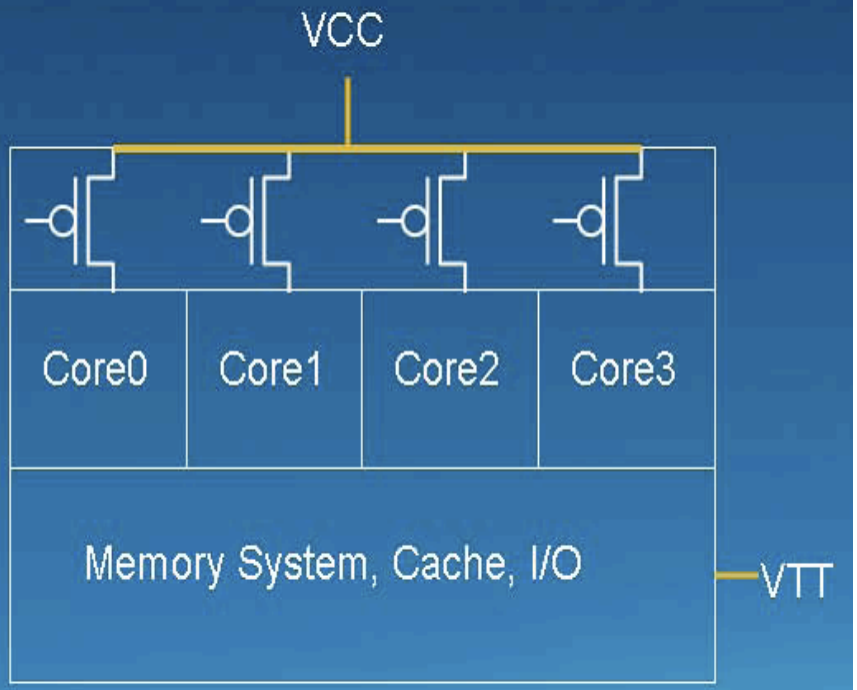


Intel Hyper-Threading Technology

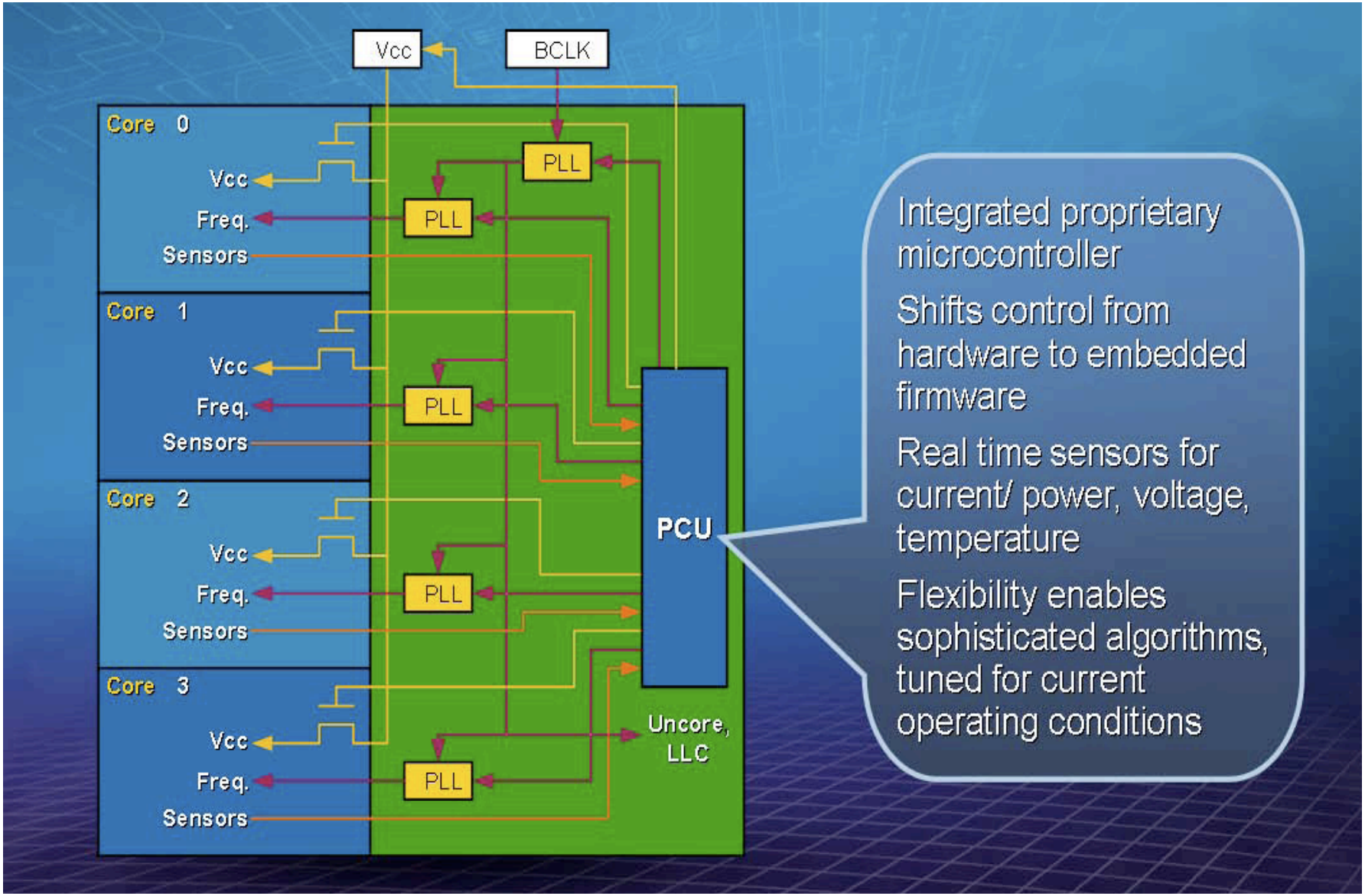


Power Power Gate

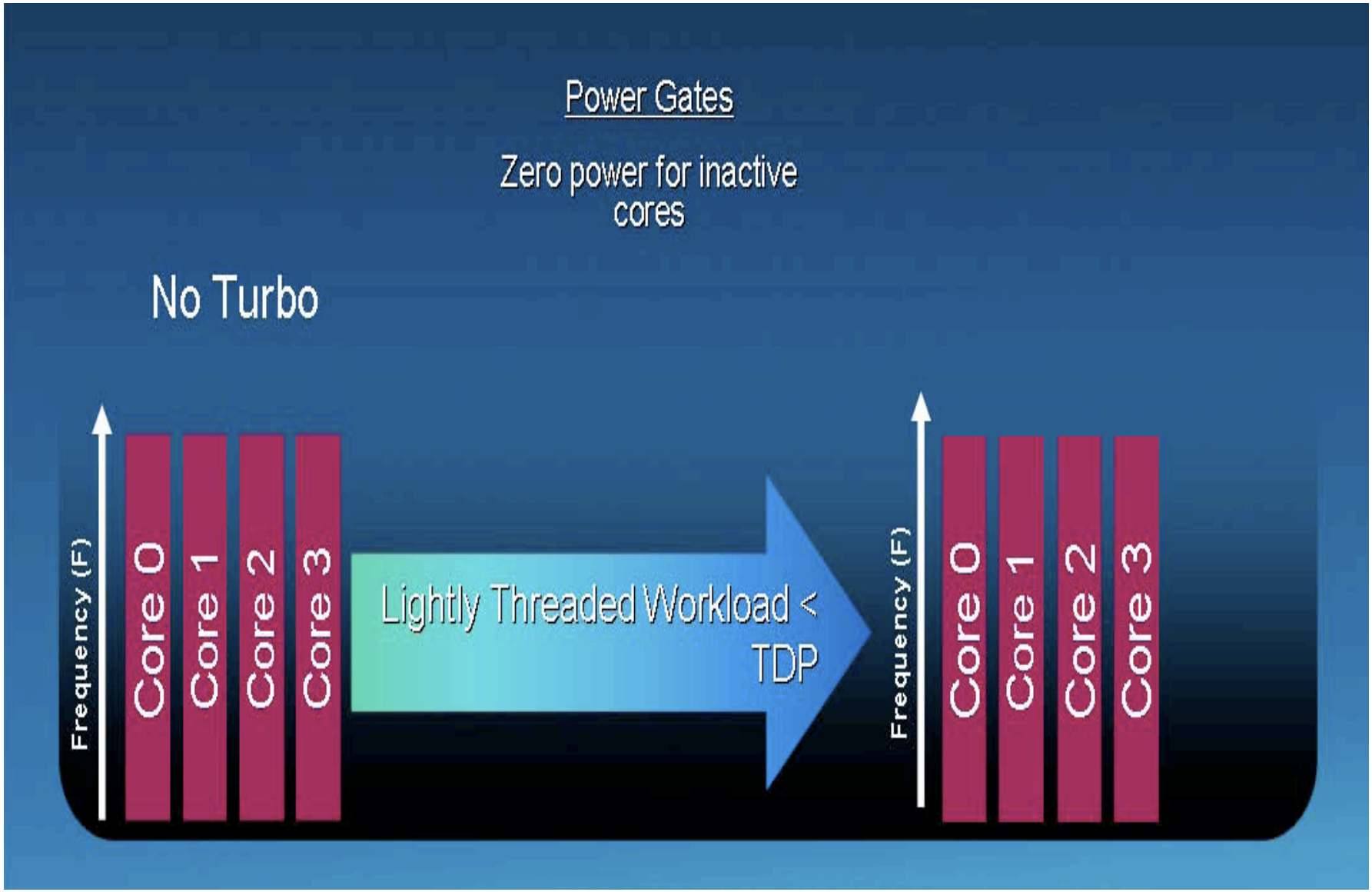
- Clock Gate
  - Exists in all modern Intel processors
  - Shuts off switching power from idle logic but leakage power remains
- Power Gate: New
  - Shuts off both switching power and leakage power
  - Enables idle cores to go to ~0 power (C6), independent of state of other cores on die
  - Completely transparent to platform and software, no incremental platform cost

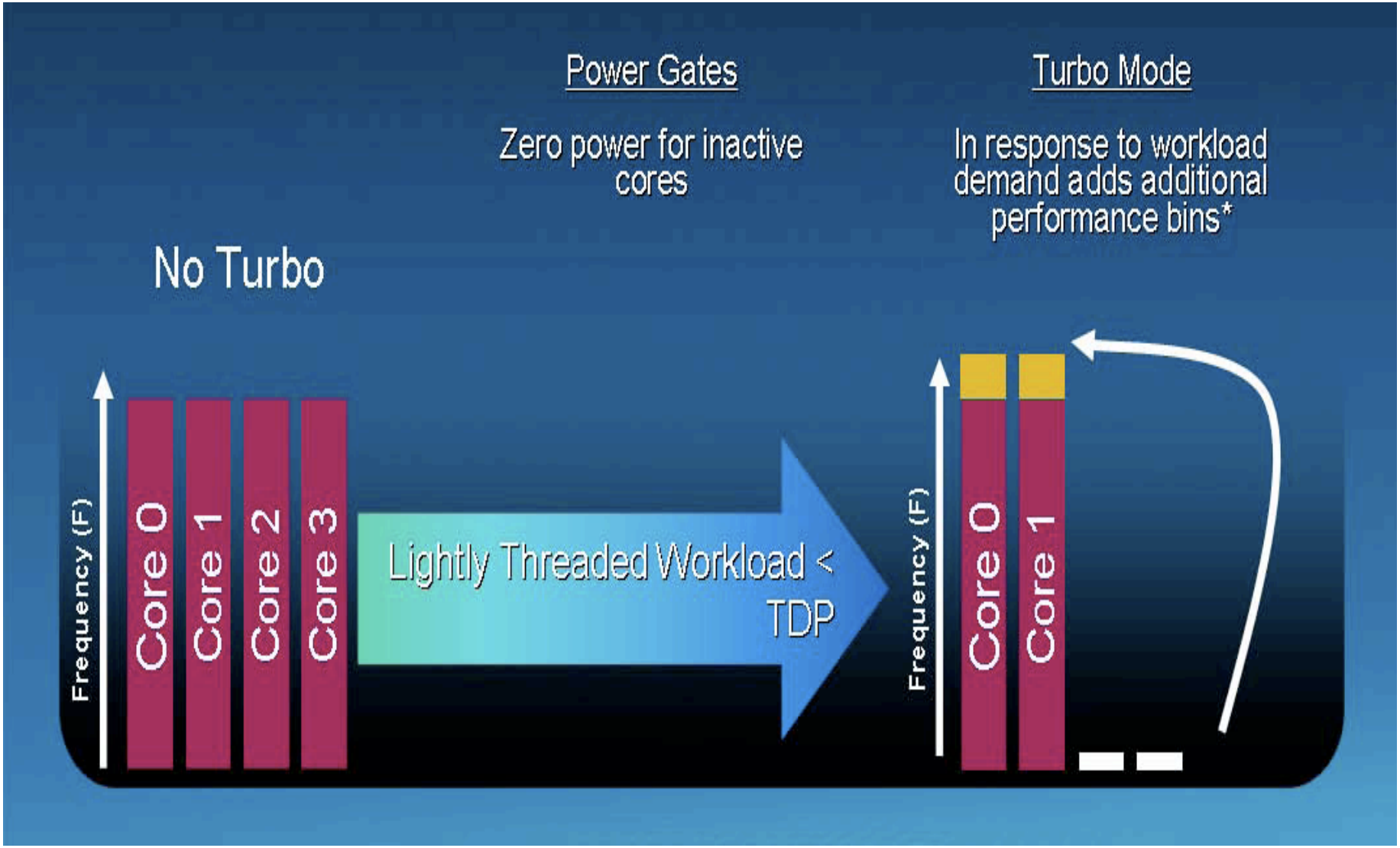


Power Management: Power Control Unit



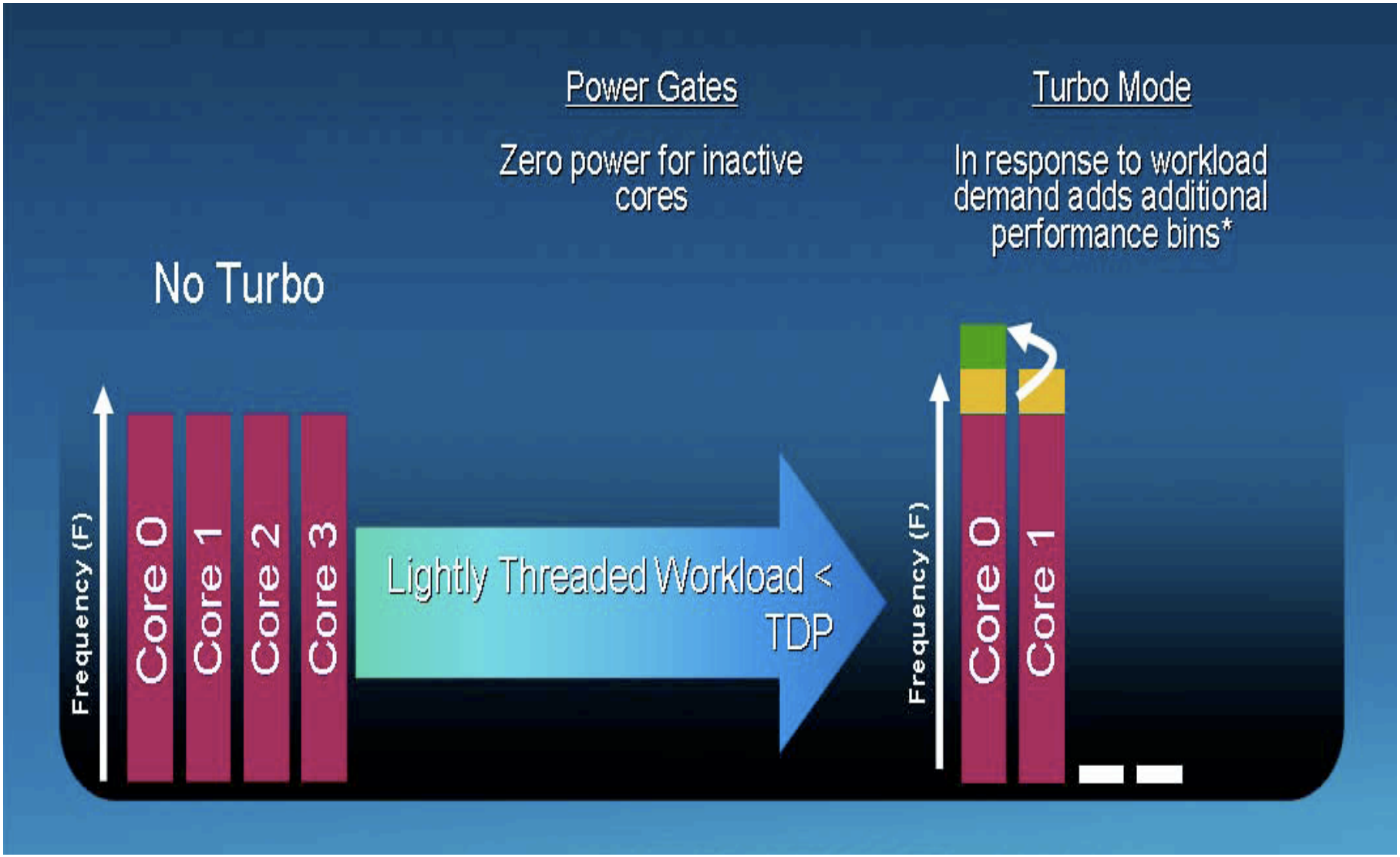
Nehalem Turbo Mode





Nehalem Turbo Mode

*** Within power and thermal constraints**



Nehalem Turbo Mode

*** Within power and thermal constraints**

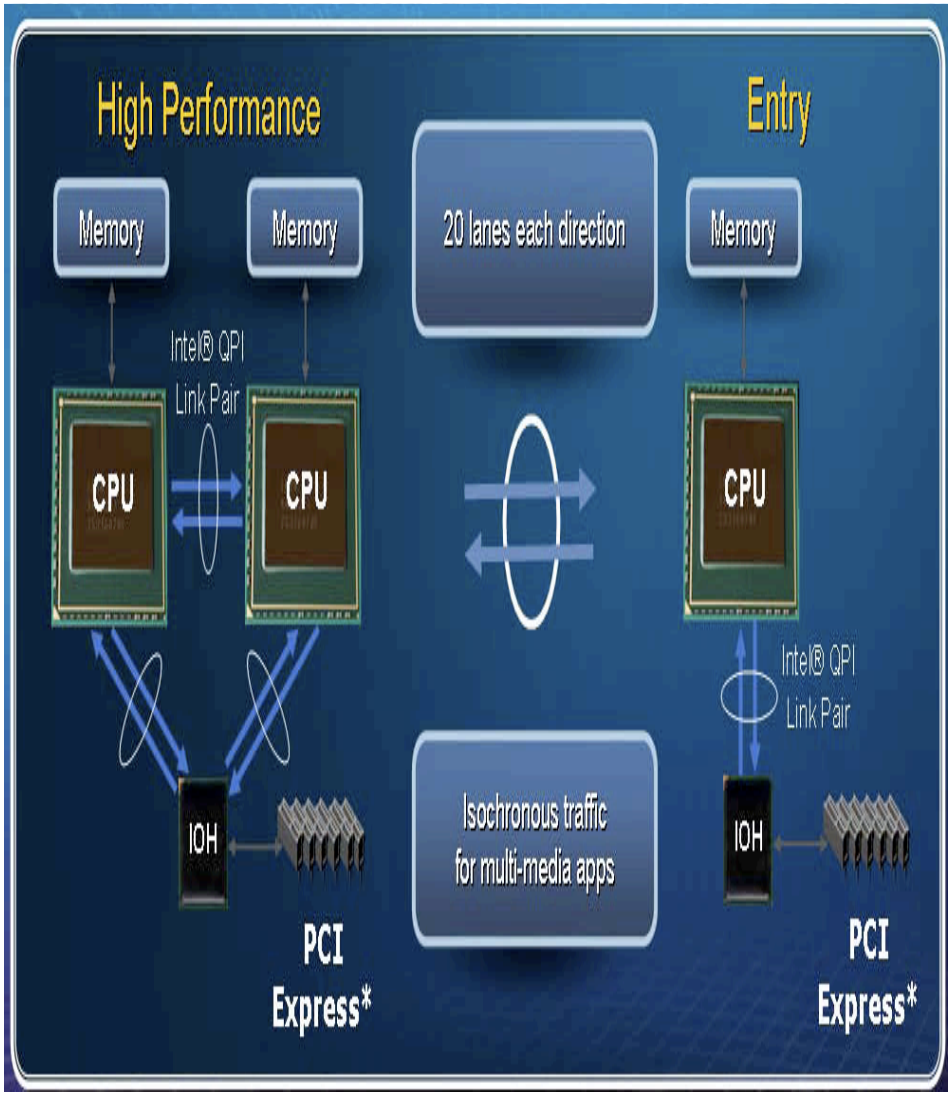
Memory Controller & DDR3 Memory

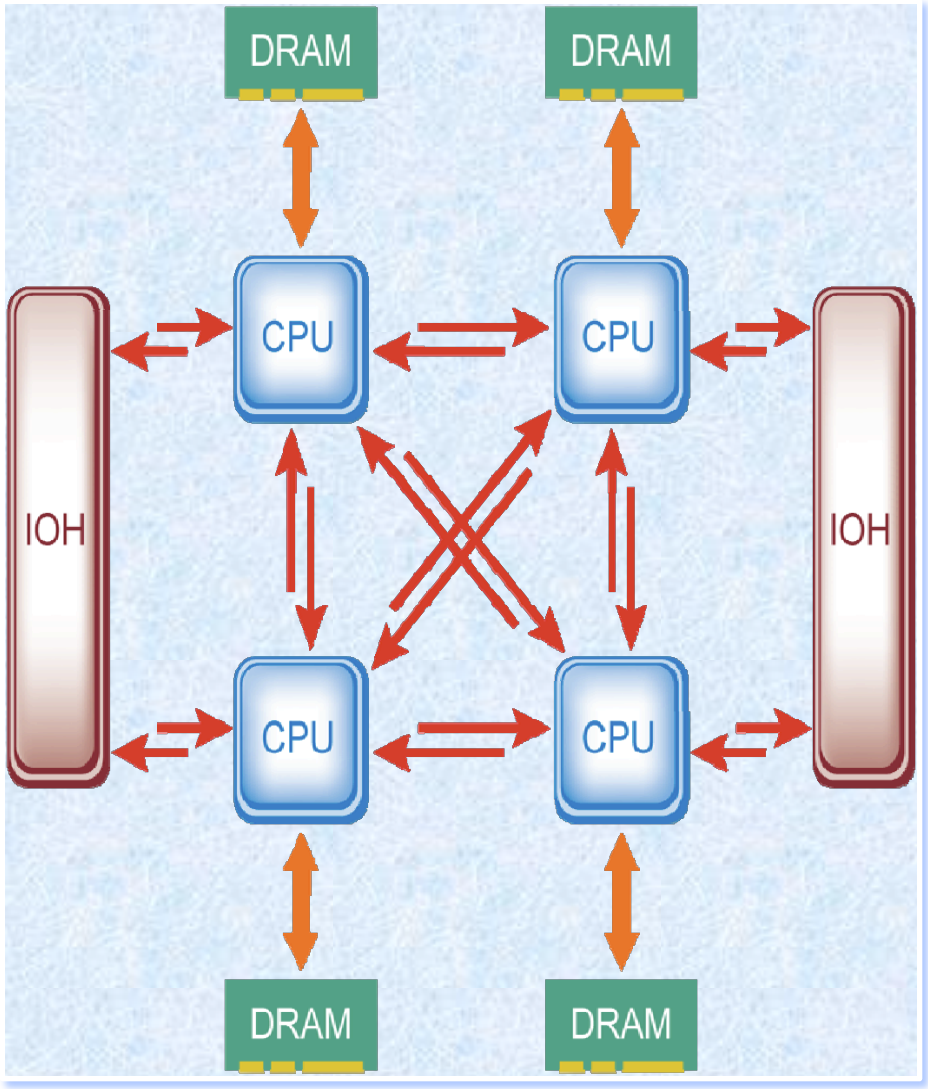
- Memory controller
  - Memory controller is a digital circuit which manages the flow of data going to and from the main memory.
  - On Nehalem, memory controller is on the microprocessor to reduce the memory latency.
- DDR3 Memory:
  - DDR3 - double-data-rate three dynamic random access memory is a random access memory interface technology.
  - Advantages of DDR3 over DDR2
    - Higher bandwidth performance
    - Slightly improved latencies
    - Higher performance at low power
    - DDR3 standard allows for chip capacities of 512 megabits to 8 gigabits

Quad Data Rate (QDR)

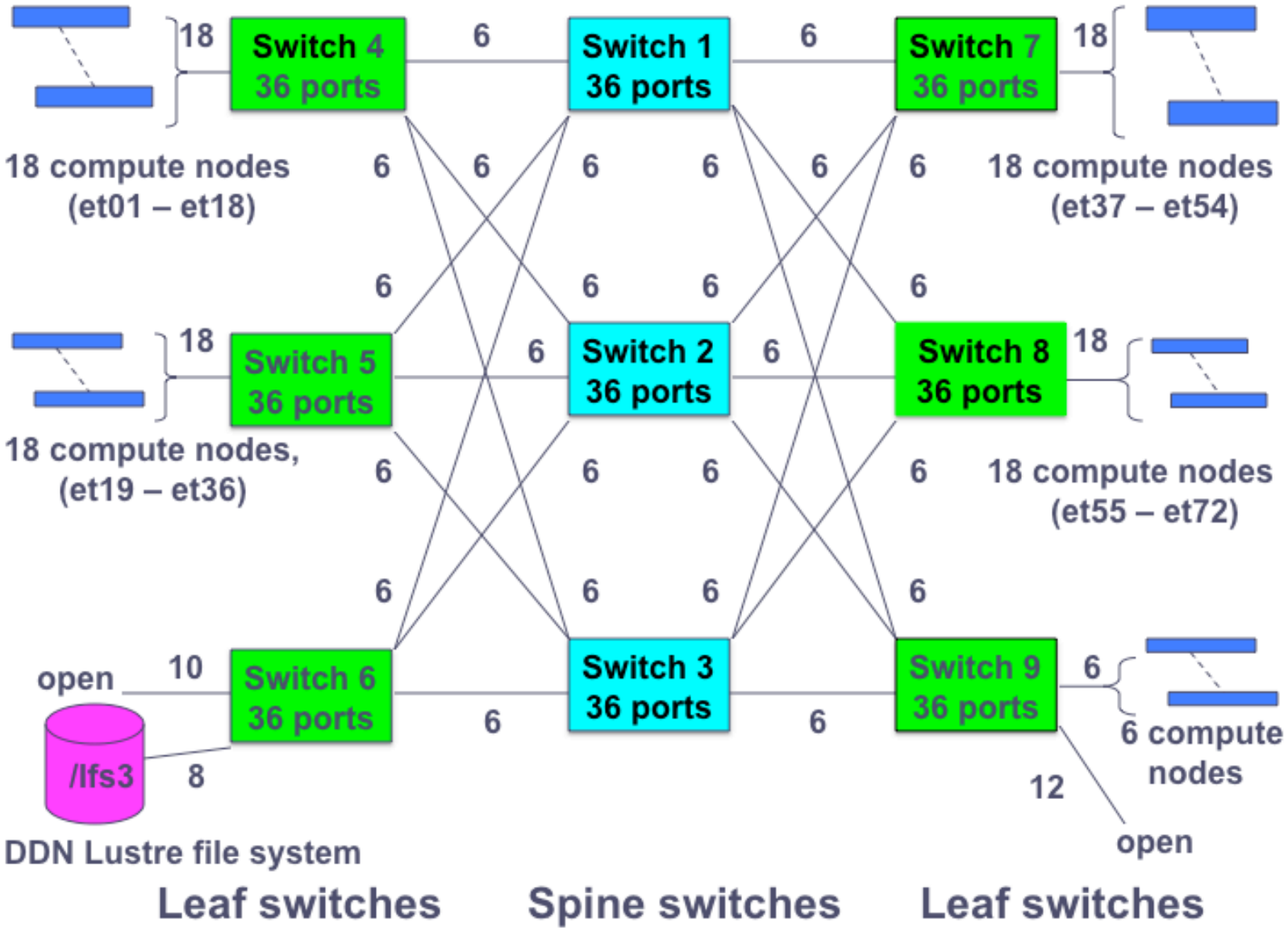
- The physical layer of InfiniBand is comprised of bidirectional links of 2.5Gb/sec. These links can be combined into 4X (10Gb/sec) and 12X (30Gb/sec) links.
- Theoretical data transfer rate is 8/10ths of the gross due to an 8/10 encoding at the physical layer.
- The InfiniBand specification also allows Double Data Rate (DDR) and Quad Data Rate (QDR) modes.
- QDR operation is clocked at quadruple the rate, allowing a 10Gb/sec signaling rate per lane.
- 4X DDR InfiniBand link has a signaling rate of 20Gb/sec, or 16 Gb/sec data rate.
- 4X QDR InfiniBand link has a signaling rate of 40Gb/sec, or 32Gb/sec data rate.

Intel Quick Path Interconnect





Intel Nehalem Cluster



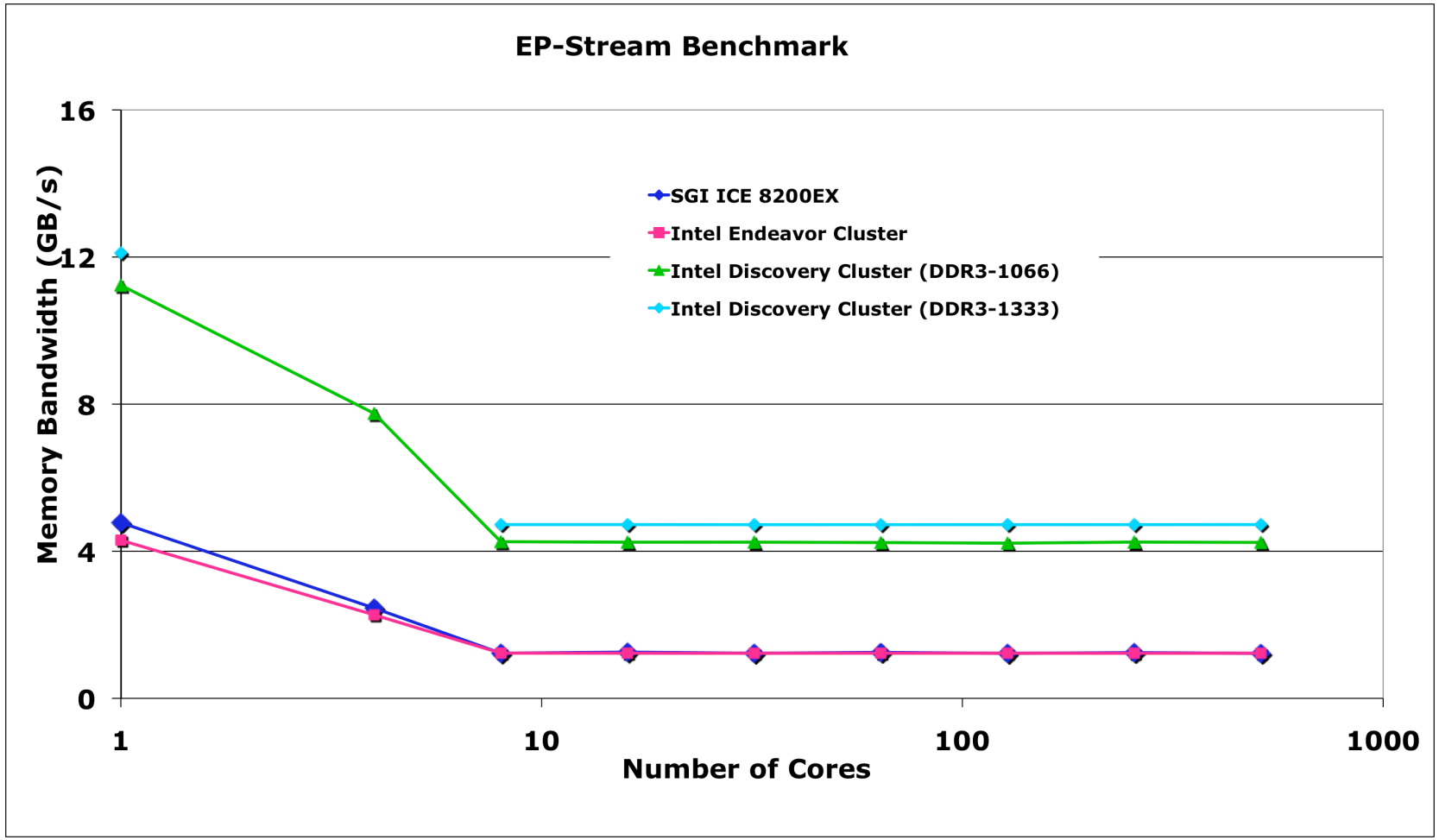
HPC Systems

- Discovery:
  - Intel cluster, Nehalem processor, 2.8 GHz, DDR3-1066/1333, 512 cores, IB QDR/DDR, Fat tree, L2$ 256KB/core, L3$ 8 MB shared and memory 3 GB/core.
- Endeavor:
  - Intel cluster, Harpertown processor, 2.8 GHz, DDR2 FB DIMM, IB DDR, Fat tree, L2$ 3 MB/core and memory 2 GB/core.
- ICE:
  - SGI Cluster, Harpertown processor, 3 GHz, DDR2 FB DIMM, IB DDR, Hypercube, L2$ 3 MB/core and memory 1 GB/core.

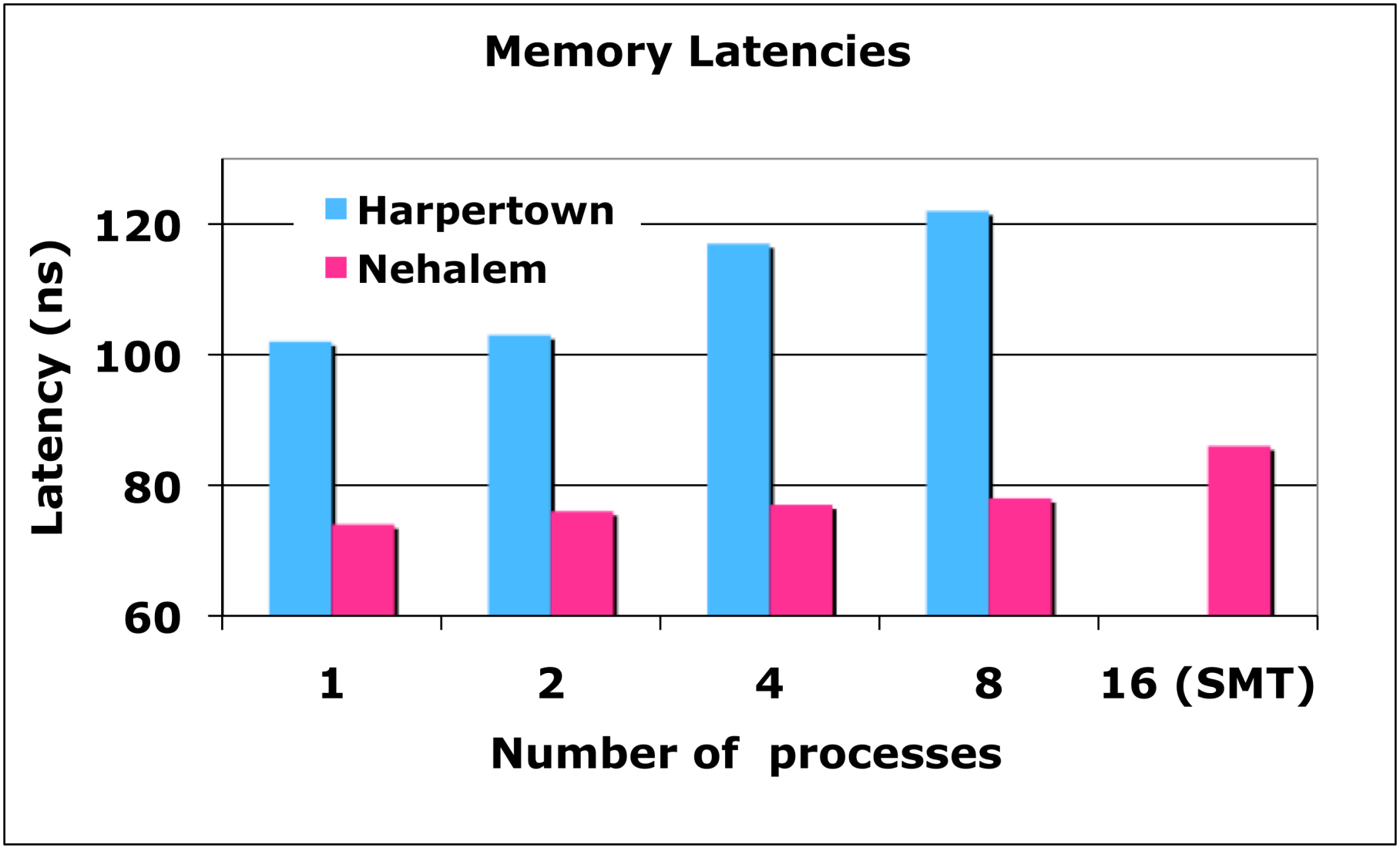
HPC Challenge Benchmarks

- Basically consists of 7 benchmarks
  - **HPL:** floating-point execution rate for solving a linear system of equations
  - **DGEMM:** floating-point execution rate of double precision real matrix-matrix multiplication
  - **STREAM:** sustainable memory bandwidth
  - **PTRANS:** transfer rate for large data arrays from memory (total network communications capacity)
  - **Random Access:** rate of random memory integer updates (GUPS)
  - **FFTE:** floating-point execution rate of double-precision complex 1D discrete FFT
  - **Latency/Bandwidth:** ping-pong, random & natural ring

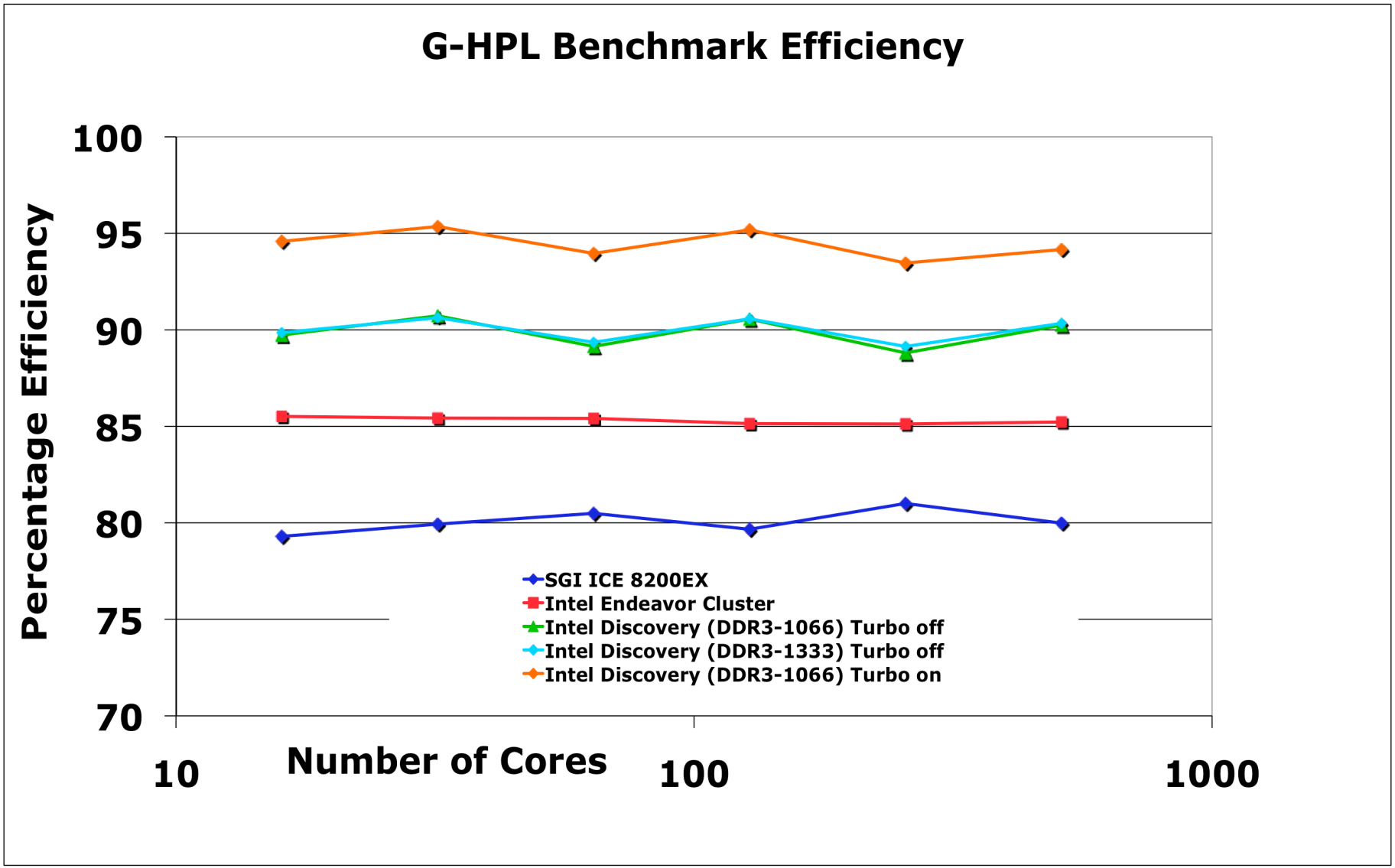
EP - STREAM


EP-Stream Benchmark
Memory Bandwidth (GB/s)
16
12
8
4
0
1
10
100
1000
Number of Cores
SGI ICE 8200EX
Intel Endeavor Cluster
Intel Discovery Cluster (DDR3-1066)
Intel Discovery Cluster (DDR3-1333)

Memory Latencies



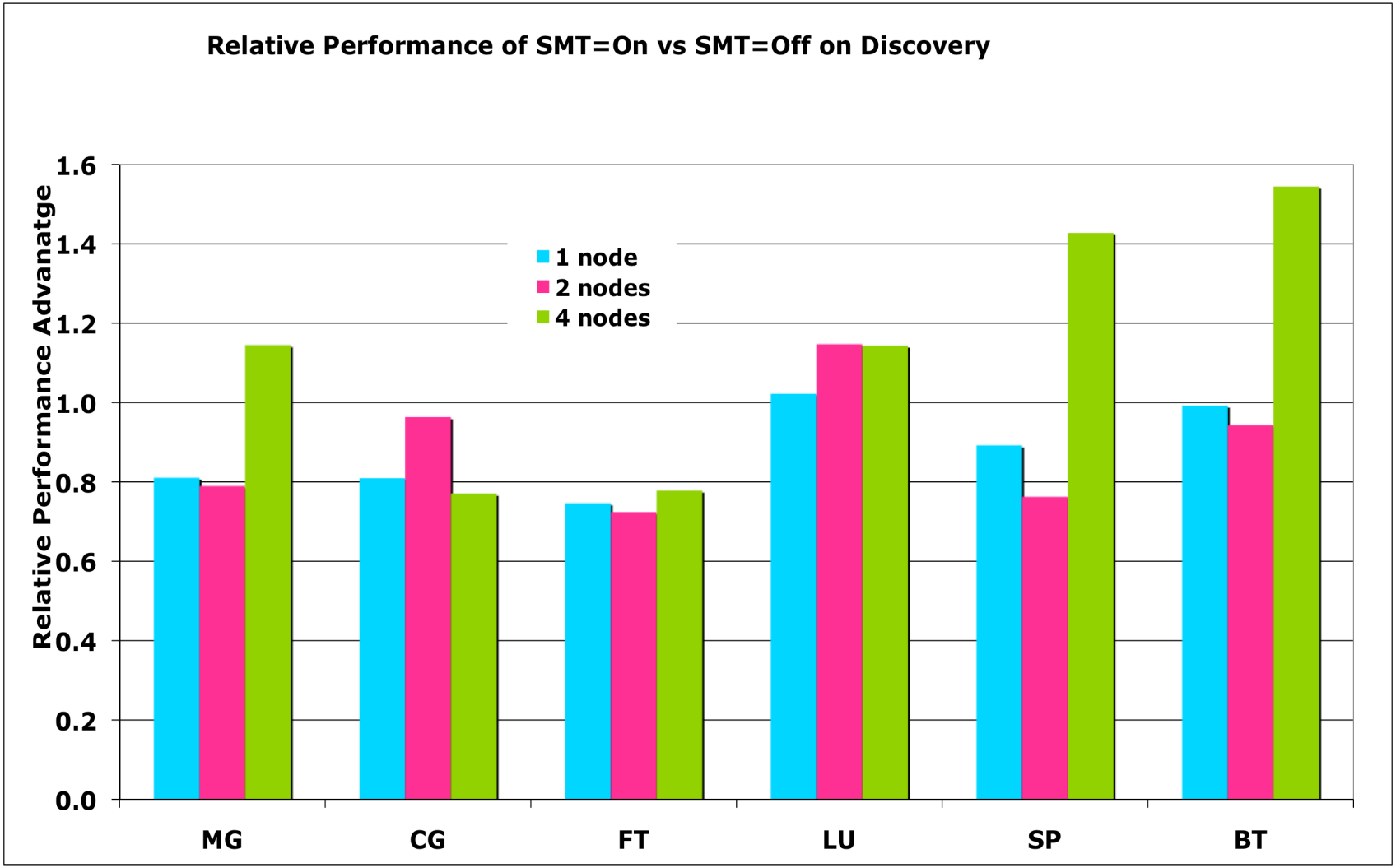
G - HPL


G-HPL Benchmark Efficiency
Percentage Efficiency
100
95
90
85
80
75
70
10
Number of Cores
100
1000
SGI ICE 8200EX
Intel Endeavor Cluster
Intel Discovery (DDR3-1066) Turbo off
Intel Discovery (DDR3-1333) Turbo off
Intel Discovery (DDR3-1066) Turbo on

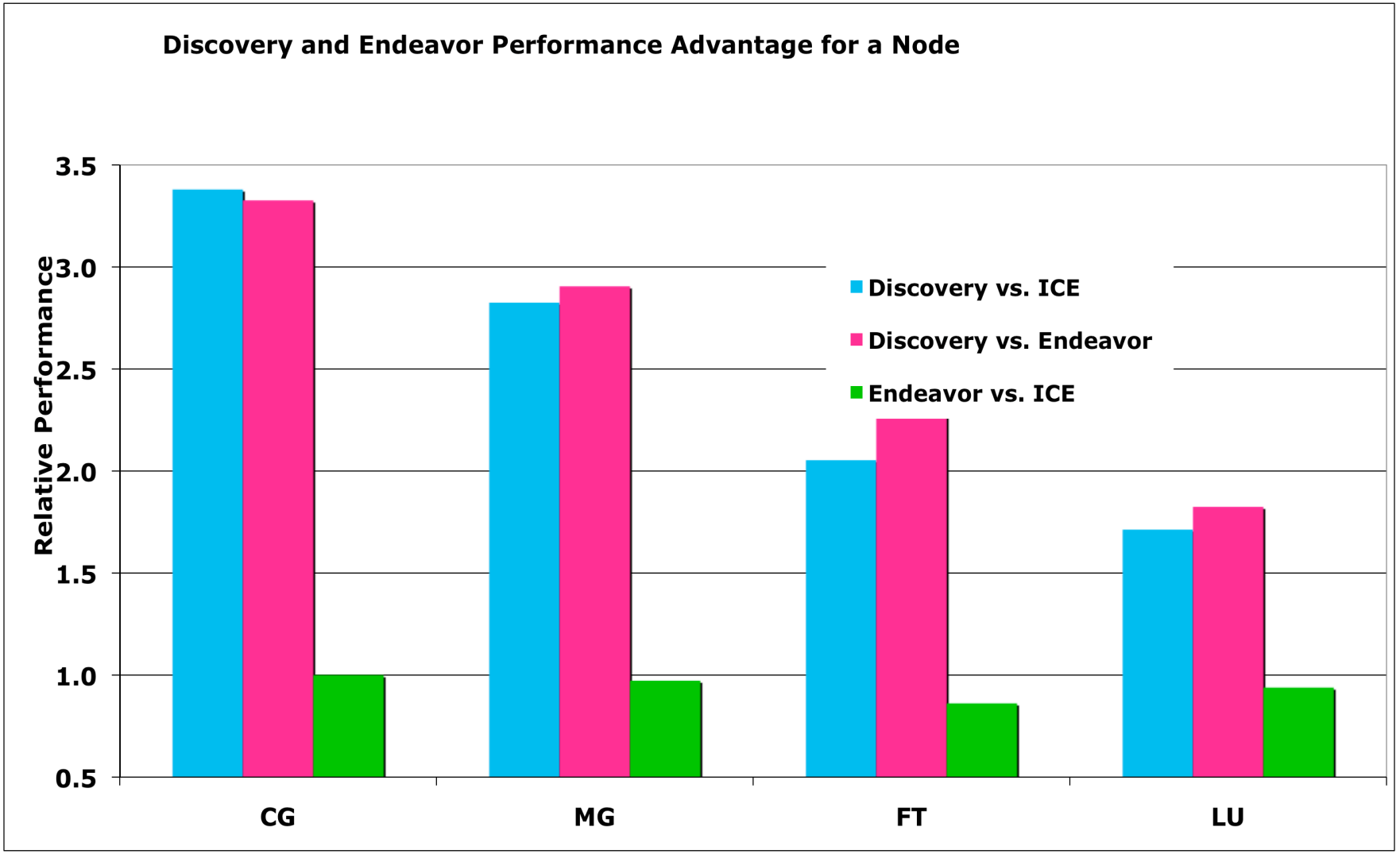
NAS Parallel Benchmarks (NPB)

- **Kernel benchmarks**
  - **MG:** multi-grid on a sequence of meshes, long- & short-distance communication, memory intensive
  - **FT:** discrete 3D FFTs, all-to-all communication
  - **IS:** integer sort, random memory access
  - **CG:** conjugate gradient, irregular memory access and communication
  - **EP:** embarrassingly parallel
- **Application benchmarks**
  - **BT:** block tri-diagonal solver
  - **SP:** scalar penta-diagonal solver
  - **LU:** lower-upper Gauss Seidel

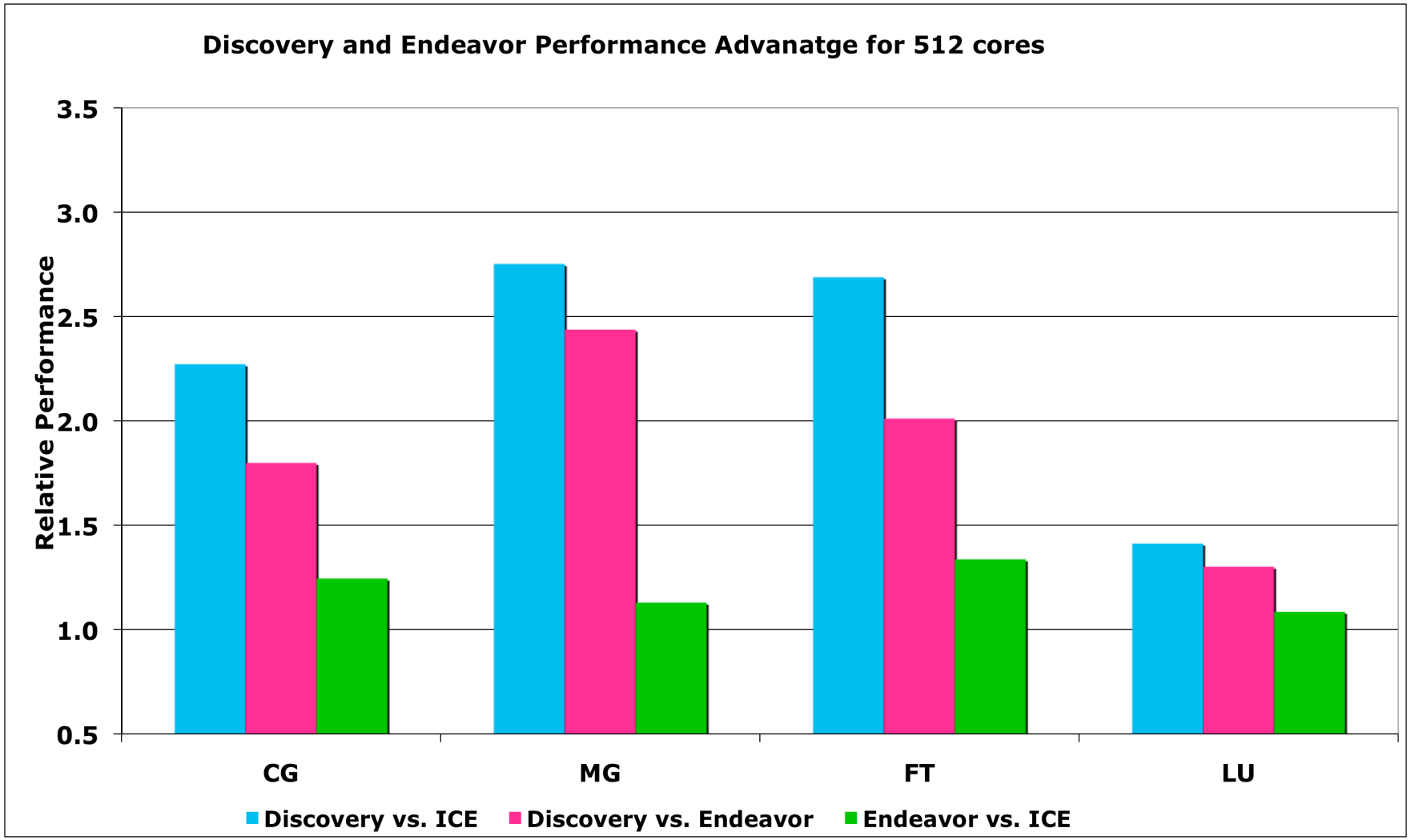
NPB: SMT=On vs. SMT=Off



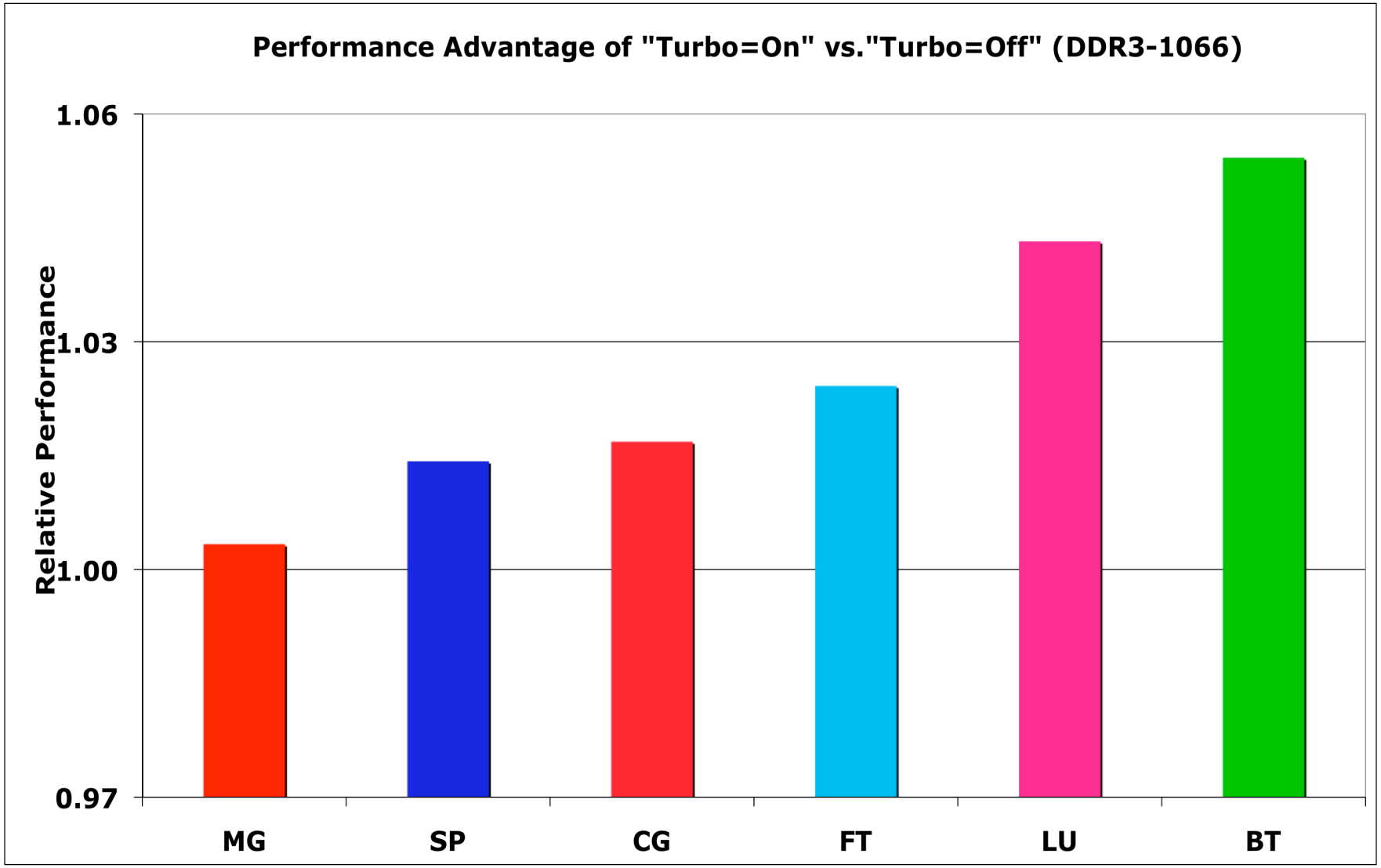
Nehalem vs. Harpertown on a Node



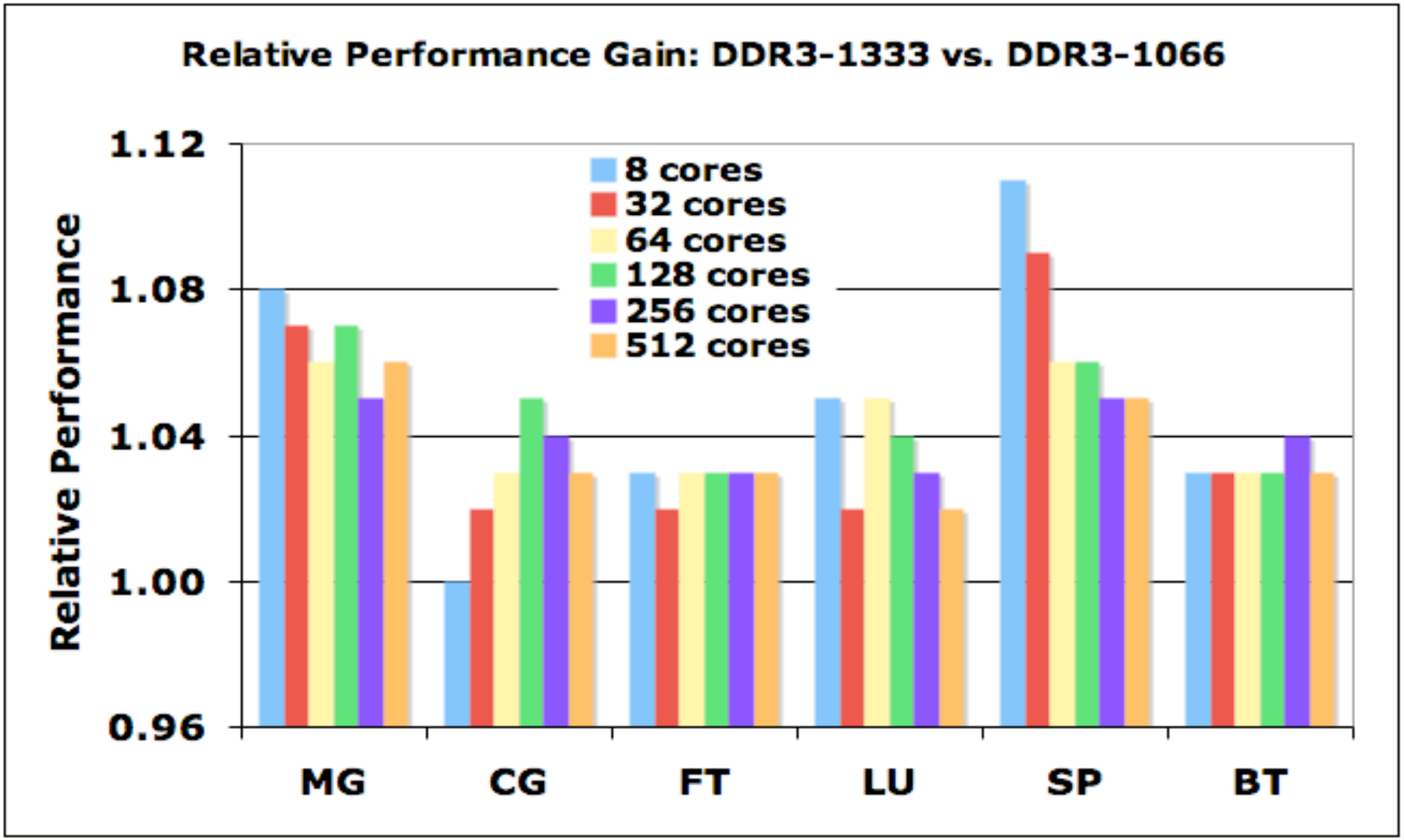
Relative Performance on NPB 512 Cores



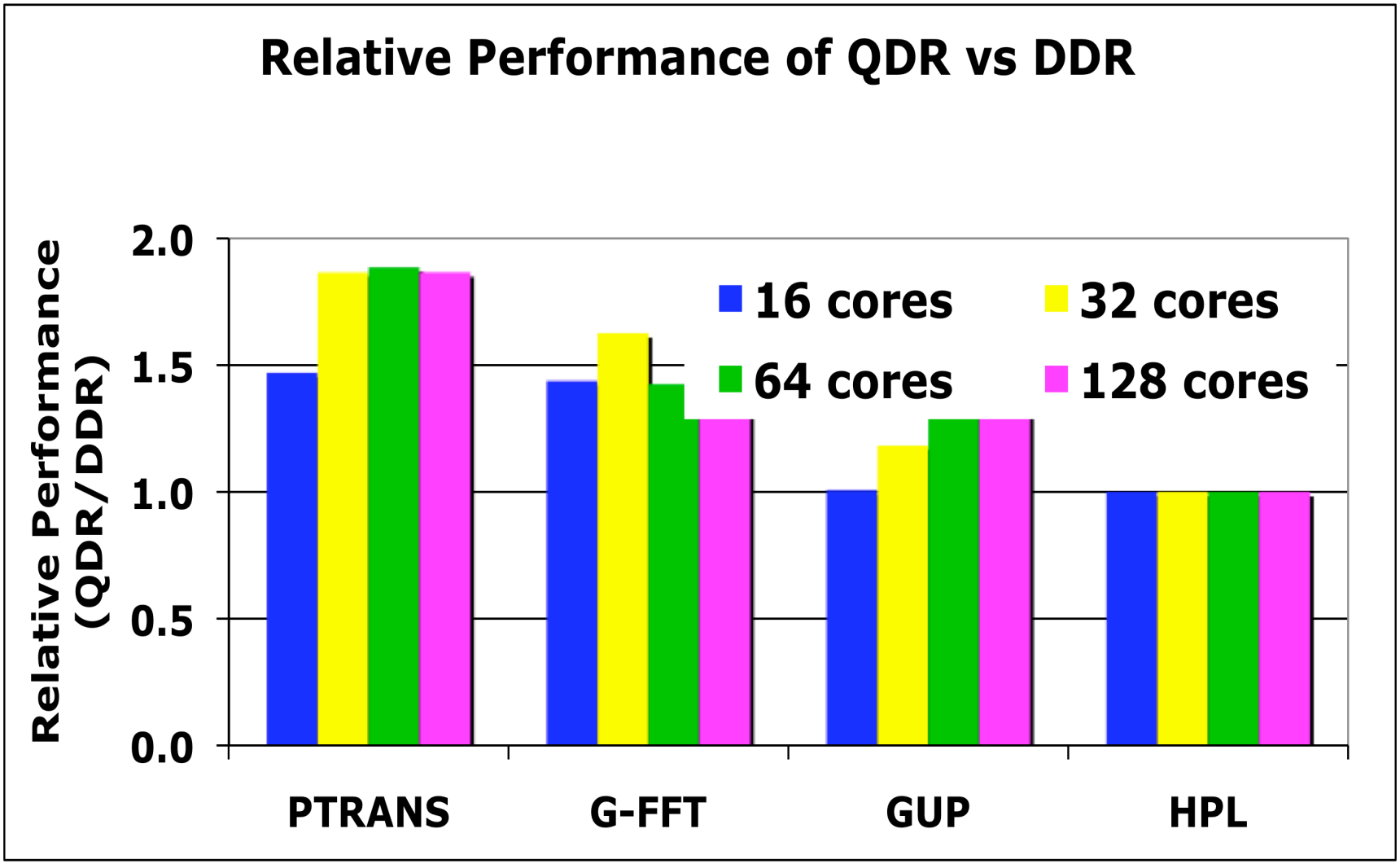
Turbo=On vs. Turbo=Off



DDR3-1333 vs. DDR3-1066



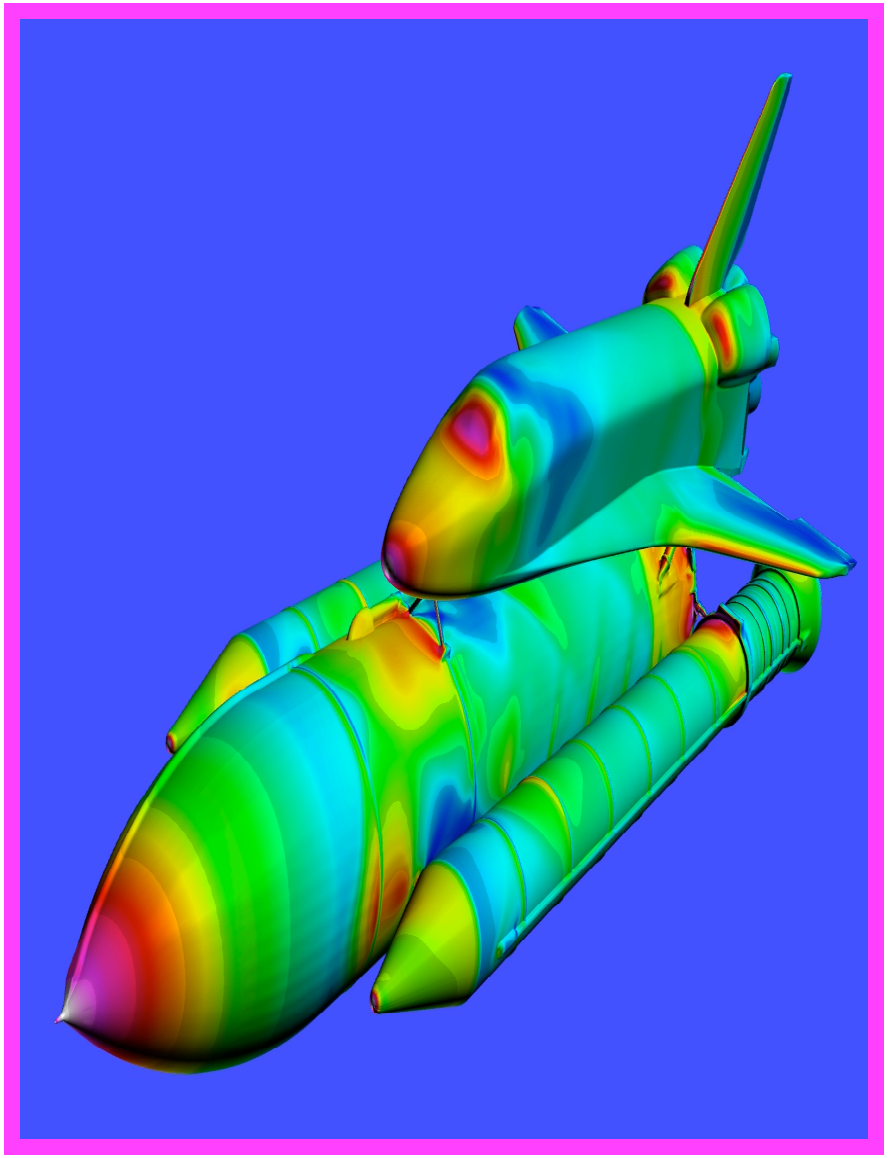
Relative HPCC Performance of QDR vs. DDR



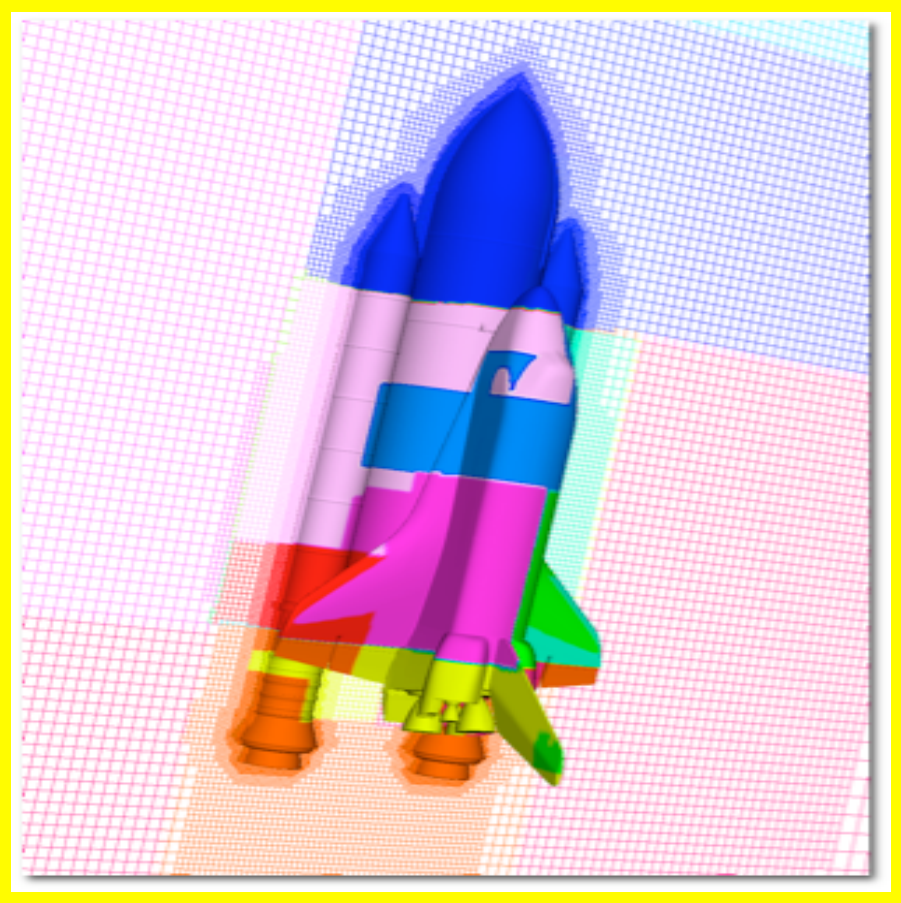
OVERFLOW-2

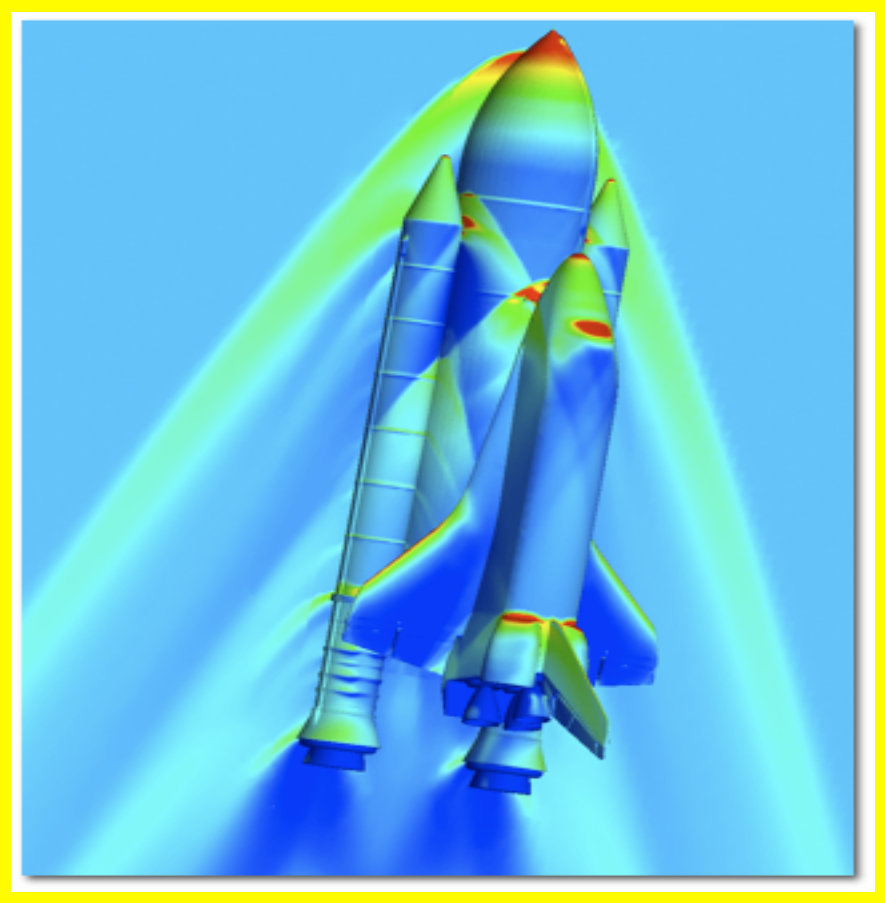
- **Code Description:**
  - Computational Fluid Dynamics code for the compressible Navier-Stokes equations (NASA Langley)
  - Finite differences in space, implicit time stepping,
  - Handles geometric complexity via overlapping grids
  - Fortran, some C
  - Has been used for multiple projects including Space Shuttle Launch Vehicle, subsonic transport aircraft
- **Code Characteristics**
  - MPI SPMD model
  - Natural parallelism with multiple overset grids
  - Further grid splitting for load balancing
  - **Memory bound**
  - Not communication intensive

Space shuttle and boosters simulated in Overflow-2



CART3D

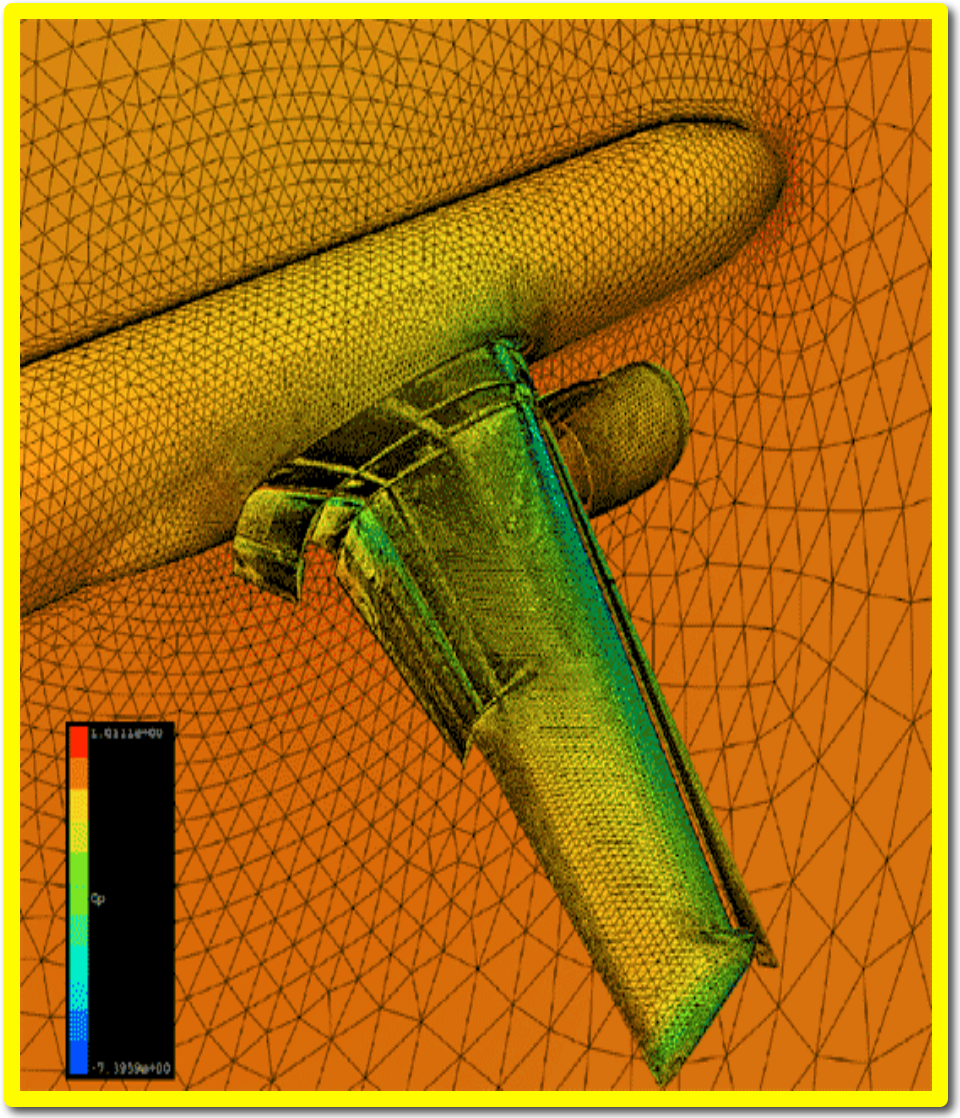




- Inviscid analysis package, Cartesian structured meshes
- Surface modeling, mesh generation, data extraction
- Space-Filling-Curve based partitioner and mesh coarsener
- Each sub-domain has own local grid hierarchy
- **CPU-intensive, memory bound for large multigrid levels**

USM3D

- USM3D is an tetrahedral based cell-centered Navier-Stokes flow solver using an unstructured meshes
- It is part of the NASA Tetrahedral Unstructured Software System (TetrUSS) suite
- Routinely used to predict aerodynamic parameters like lift, drag, and detailed airflow about candidate aircraft and aerospace vehicle designs
- **The model is memory bound and latency bound for large number of cores**

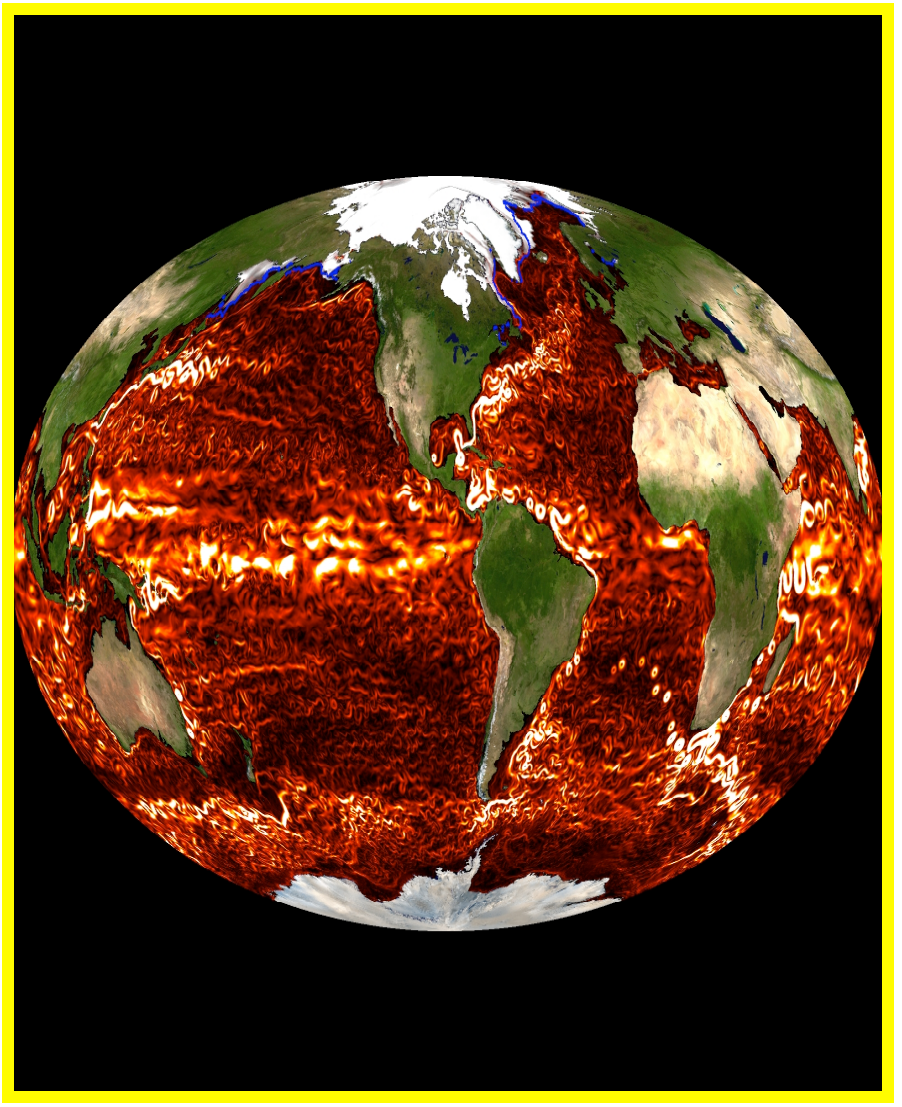


108M Tetrahedral grid used in Boeing 777 Simulation

ECCO

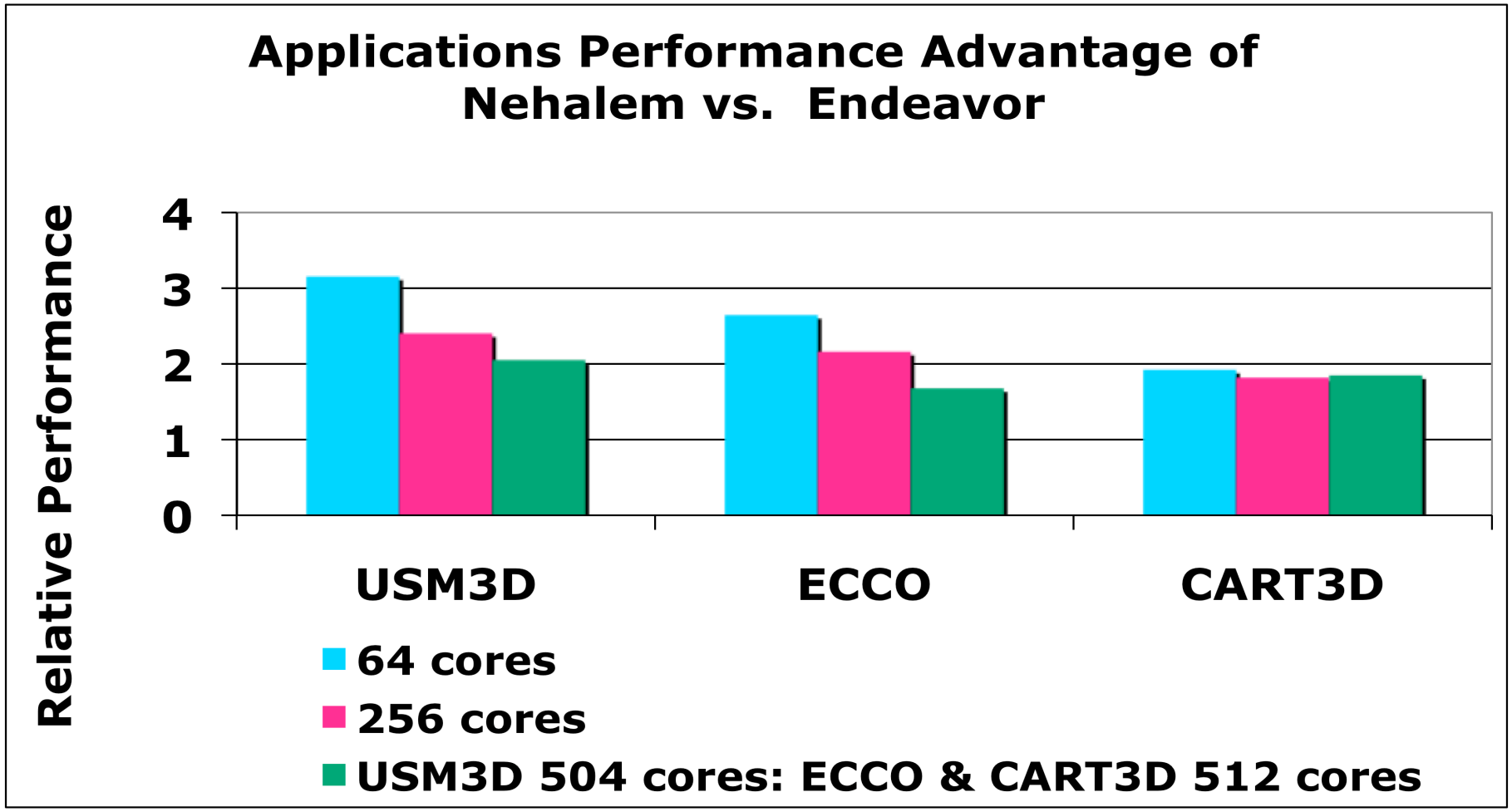
Magnitudes of velocity across the globe using ECCO

- Estimating the Circulation and Climate of the Ocean (ECCO) is a global ocean circulation model solving the hydrostatically approximated time dependent Reynolds averaged Navier-Stokes equations in 3D
- Utilizes a finite volume discretization of the equations of motion that is 2nd order accurate in time and space
- **Memory bound.**
- **Scaling is ultimately latency bound for large number of cores**
- **Significant amount of I/O**

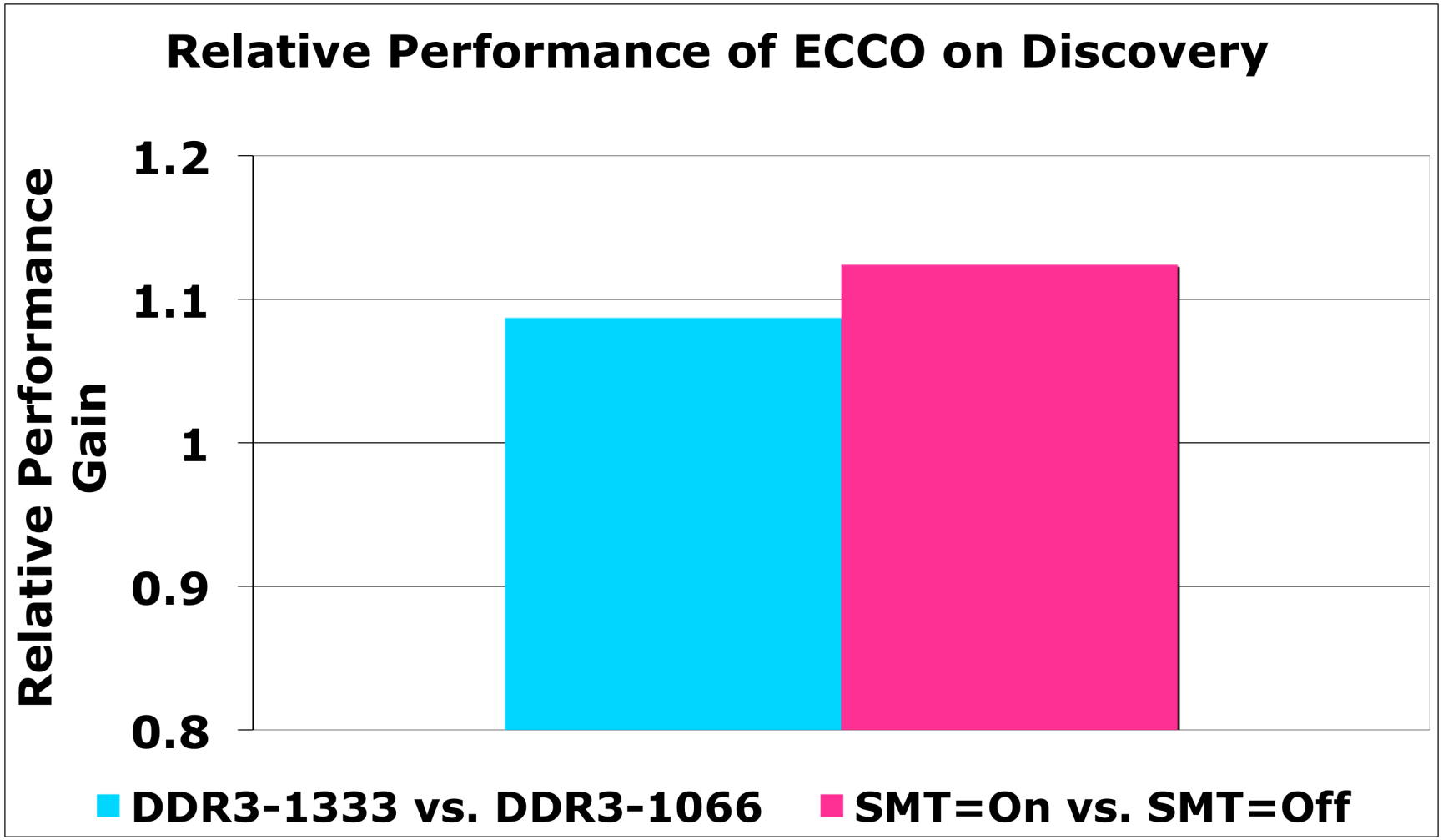


1/4th degree global simulation

Applications Performance Advantage



Relative Performance of ECCO Application



Conclusions

- A single process on a Nehalem system has more available bandwidth than the accumulated bandwidth of 8 processes on the Harpertown system
- The observed memory latency on the Nehalem system using all 16 SMTs simultaneously is lower than that of Harpertown using a single core
- SMT technology improves the performances of some applications but for HPC applications it is not universal. Experimentation is recommended.
- Turbo mode helps compute bound applications.
- Performance advantage of real world applications over Harpertown is between 1.9 and 2.1 for higher core counts and 3.2 for lower ones.
- QDR interconnects helps only bandwidth bound applications.